High-Performance | Multi-Purpose | Bioinformatics

# Lenovo GOAST Bioinformatics Solution

A Genomics and Bioinformatics Optimized System

Smarter technology for all

Lenovo

## Highlights

**Extremely fast analytics**
Bioinformatics optimized hardware runs sequential workflows faster: e.g. processes a 30x whole genome in ~18 minutes and a 50x whole exome in ~30 seconds

**Multi-purpose Bioinformatics use**
Leverage BOSS's high-core, fast I/O, and high-memory specs to run any Bioinformatics or HPC tools or scripts

**Increases lab productivity**
Faster time to insight: e.g. up to ~11K whole genomes/node/year

**Cost effective**
Up to 50% less than boutique solutions relying on GPUs or FPGAs without additional licensing fees. A single GOAST server can replace up to 50 standard nodes

**Scalable**
Deploy as a single-node appliance or as a cluster and grow linearly with flexibility

Bioinformatics is the computational analysis of biomedical data powering research all the way from basic biology, to medicine, to drug discovery, to agriculture, and more. The deluge of bioinformatics data generated in the last two decades has prompted researchers to gradually shift their workloads from desktop to cluster to supercomputer analytics. Yet to date, even at cluster and supercomputer speeds, large-scale bioinformatics still faces long execution times on massive volumes of data, which delays "time to answer".

In turn, lengthy analyses severely impact academics' ability to obtain grants and publish papers and companies' growth and profits. The only high-performance solutions to mitigate these challenges are single-purpose boutique solutions requiring expensive specialty hardware and substantial licensing fees. These single-purpose solutions require organizations to purchase multiple architectures to support other types of research in their datacenter since bioinformatics analytics does not exist in a vacuum. Even those organizations working in a single subfield of Bioinformatics (e.g. in genomics) find that the single-purpose boutique solutions are not enough.

For example, those performing secondary genomics analytics find that they also need computational resources to gather, select, store, manage, transform, and describe their data, in both primary and tertiary downstream analyses. General-purpose datacenters worldwide feel this need even more acutely since the Omics (Genomics, transcriptomics, proteomics) are only a fraction of the users they must serve. Therefore, Lenovo developed the GOAST system, a Genomics and Bioinformatics Optimized platform.

## Extremely Fast Bioinformatics Analytics

Lenovo GOAST is a multi-purpose system specifically engineered to meet the demands of bioinformatics workloads. GOAST leverages an architecture of carefully selected hardware tuned to accelerate bioinformatics performance. Lenovo GOAST's high-core, fast I/O, and high-memory specs (Table 1) excel at running the massively parallel applications and sequential workflows common in Bioinformatics, including multi-omics (genomics, transcriptomics, proteomics) applications. GOAST accelerates mapping and whole-genome sequencing (WGS) variant calling analytics from days to minutes—a process that today in many datacenters around the world takes 40-150 hrs. runs in just ~18 minutes in GOAST systems (Table 2). GOAST Plus runs the total end-to-end germline workflow execution in 48 min. for the median (or 54 min. for the last) sample in a batch run when accounting for all steps and job queuing. The GOAST Base system, a price-performance option (Table 3), runs mapping/alignment and variant calling in ~58 minutes or ~3.0 hr. for the end-to-end workflow. GOAST also processes whole-exome sequencing (WES) samples at 50x coverage in as little as ~30 seconds or ~1.5 minutes in the Plus and Base configurations, respectively.

**Table 1:** GOAST reference architectures for bioinformatics

| | GOAST Base | GOAST Plus |
|---|---|---|
| **Processor** | 2x Intel 6248R CPUs | 8x Intel 8280 CPUs |
| **Memory** | 384GB RAM, 12x 32GB/2933MHz DIMMs | 1.5TB RAM, 48x 32GB/2933MHz DIMMs |
| **Storage** | Min. 2TB SAS SSD | Min. 4TB NVMe |

*The GOAST reference architecture is fully customizable

**Table 2:** Execution of the median sample when running a batch of 30x NA12878 WGS samples on the GATK v.4.2 Germline variant calling pipeline on a GOAST Plus system

| Step | Min./sample |
|---|---|
| SamToFastqAndBwaMemAndMba | **11.3** |
| MarkDuplicates | **3.6** |
| SortSampleBam | **9.0** |
| BaseRecalibrator | **2.5** |
| GatherBamFiles | **1.4** |
| ApplyBQSR | **1.5** |
| HaplotypeCaller | **6.2** |
| MergeVCFs | **1.4** |
| End-to-end Workflow | **47.9** |
| Mapping + Calling | **17.6** |

**Table 3:** Execution of a the median sample when running a batch of 30x NA12878 WGS samples on the GATK v.4.2 Germline variant calling pipeline on a GOAST Base server

| Step | Min./sample |
|---|---|
| SamToFastqAndBwaMemAndMba | **39.6** |
| MarkDuplicates | **8.8** |
| SortSampleBam | **19.7** |
| BaseRecalibrator | **6.6** |
| GatherBamFiles | **6.2** |
| ApplyBQSR | **5.5** |
| HaplotypeCaller | **18.2** |
| MergeVCFs | **6.9** |
| End-to-end Workflow | **181.3** |
| Mapping + Calling | **57.8** |

## Increases lab productivity

Accelerated execution speeds mean you get to process more samples, find answers faster, and generate breakthroughs that much sooner. GOAST outperforms any other competing CPU-based (and even the FPGA- and GPU-based) systems because we tune our systems to meet the requirements of bioinformatics pipelines running in-node workloads rather than those assumed in traditional HPC workloads. The result is the ability to run software pipelines in higher throughputs. Higher throughput capacity means batches of samples analyzed in less time. **(Table 4)**.

**Table 4:** Lab Productivity for Omics expected on a single GOAST system given the performance in Table 2 and Table 3

| Expected Lab Productivity | 30x WGS Samples processed (n) | | 50x WES Samples processed (n) | |
|---|---|---|---|---|
| | GOAST Plus | GOAST Base | GOAST Plus | GOAST Base |
| **WGS/day/node** | 30 | 8 | 999 | 279 |
| **WGS/year/node** | 10,982 | 2899 | 364,763 | 101,713 |

## Multi-purpose Bioinformatics use

GOAST is a high-performance system for multi-purpose Bioinformatics use. The system comes preloaded with Omics tools to get you up and running on day one or it can be fully customized with the Bioinformatics tools of your choice.

**For multi-omics analytics:** Lenovo pre-installs the tools and other dependencies in Table 5 necessary to run the Broad Institute's GATK Best Practices for Germline SNP and Indel discovery. The Omics methods in Table 5 provide mapping and alignment, sorting, duplicate marking, and haplotype variant calling algorithms. Lenovo GOAST also provides pre-configured scripts to allow you to run (submit, monitor, manage) samples on the Germline workflow optimally on Lenovo hardware.

**The tools pre-installed in Table 5 will allow you to run the following analyses with your own scripts:**

- Germline SNP and Indel
- Germline joint genotyping
- Germline copy number variant (CNV)
- Somatic short variant discovery (SNVs + Indels)
- RNAseq (SNPs + Indels)

**For other Bioinformatics:** Install any tools of your choice (see examples in Table 6) on GOAST systems or talk to our team about pre-installing your software pipeline of choice.

Smarter technology for all

Lenovo

**Table 5:** Genomics analytics software and other dependencies pre-installed by GOAST 2.0

| Software | Version |
|---|---|
| GATK | **4.2.0.0** |
| BWA | **0.7.17** |
| Samtools | **1.11** |
| Picard Tools | **2.25.0** |
| JSONPP | **1.3.0** |
| Slurm | **20.02.5** |
| java | **java-1.8.0-openjdk-devel** |
| OS | **CentOS 7.9** |

*GOAST systems currently use Cromwell as the workflow manager, Slurm as the job scheduling system, and MySQL to manage and collect workflow metadata.

## Cost effective

GOAST leverages an optimized CPU-based architecture thus it requires no FPGAs or GPUs of any kind for acceleration. A CPU-based infrastructure and open source tools mean costs 50% or lower than boutique solutions relying on FPGAs or GPUs and no licensing fees. The Lenovo Bioinformatics R&D group continually tests new bioinformatics pipelines and releases to its customers hardware-tuned versions of standardized workflows such as the Broad Institute's GATK Best Practices at no cost. In addition, GOAST solutions can reduce investments needed to support large-scale projects since a single GOAST Plus server can replace up to 50 standard nodes, reducing hardware, maintenance costs, and other expenses, including power consumption and cooling.

**Table 6:** Examples of the multi-purpose uses for the GOAST systems

| Field | Bioinformatics Method | Purpose |
|---|---|---|
| **Genomics** | Genome assembly<br>Variant calling | Decipher instructions in DNA<br>Genotype to phenotype |
| **Transcriptions** | RNAseq | Decipher RNA instructions |
| **Proteomics** | Gene/Protein alignment Protein/DNA Similarity searching<br>3D Structural prediction<br>Molecule interactions | Decipher Proteins<br>Drug discovery<br>Disease diagnosis<br>Biomaker identification |
| **Cheminformatics** | Molecular Dyamics<br>CryoEM | Drug design |
| **General Bioinformatics** | Data wrangling<br>Data mining<br>Biostatistics<br>Phylogenetics<br>Custom scripts | Everything else |

## Easy-to-use

For omics analytics, we preinstall the opensource tools in Table 5. At no cost to the user, we also provide the GOAST utility—a set of wrapper scripts to submit, monitor, manage, and simplify running GATK pipelines at the command-line. You can run any of the pre-installed omics pipelines either on the bare metal installation or as a Docker container. GOAST systems can be easily integrated into new or existing clusters and can be fully customized from an architecture, system, or software perspective.

## Scalable

The performance of Lenovo GOAST scales linearly from single-node appliance to cluster implementation to serve the needs of labs of all sizes, from small research groups, to commercial labs, and to national population-level projects. This includes transitioning from WES to WGS, undertaking a new project with greater scope and complexity, and expanding both data and users. Scale linearly simply by adding compute and storage building blocks as needed.

## Which GOAST Configuration is right for me?

| About you | GOAST Base | GOAST Plus |
|---|---|---|
| **Your top priority** | **Budget** | **Fastest performance** |
| **WGS mapping + variant calling performance you seek** | 58 minutes/WGS | 18 minutes/WGS |
| **WES mapping + variant calling performance you seek** | 2 minutes/WES | 30 seconds/WES |
| **The number of WGS you need to process daily per node** | ~8 WGS/day/node | Min. 4TB NVMe |
| **The number of WGS you need to process annually per node** | ~3K WGS/yr./node | ~11K WGS/yr./node |
| **You regularly process** | 1-10 WGS samples at a time per node | >10 WGS samples at a time per node |
| **If you have to process >10K WGS/yr. and...** | Don't mind managing a multi-node solution | Prefer simplicity of managing fewer nodes |
| **Regarding points of failure** | Want to avoid a single point of failure | Not overly concerned with single points of failure |
| **Best option for...** | General omics and bioinformatics | High throughput (batches of samples) or high-memory (e.g. de novo assembly) workloads |

## To Learn More

**Visit us at**:

- GOAST case study on Deep sequencing and De novo assembly at the University of Delhi
- GOAST in Coronavirus research
- GOAST in Precision Medicine
- Animated GOAST video
- GOAST site

Smarter technology for all

Lenovo

For questions reach out to: Mileidy Giraldo, Ph.D., Global lead, Life Sciences, Lenovo HPC & AI. **mgiraldo@lenovo.com**